Enriching PubMed Related Article Search with Sentence Level Co-citations

Nam Tran[1], Pedro Alves[2], Shuangge Ma[3] and Michael Krauthammer[1,2*]
[1] Department of Pathology, Yale School of Medicine,
[2] Program for Computational Biology and Bioinformatics,
[3] Department of Epidemiology and Public Health
Yale University, New Haven, CT 06520, USA

Abstract

PubMed related article links identify closely related articles and enhance our ability to navigate the biomedical literature. They are derived by calculating the word similarity between two articles, relating articles with overlapping word content. In this paper, we propose to enrich PubMed with a new type of related article link based on citations within a single sentence (i.e. *sentence level co-citations* or SLCs). Using different similarity metrics, we demonstrated that articles linked by SLCs are highly related. We also showed that only half of SLCs are found among PubMed related article links. Additionally, we discuss how the citing sentence of an SLC explains the connection between two articles.

Introduction

PubMed query and related article search are among the most useful and effective information retrieval tools for biologists and health care professionals. It has been demonstrated that the related article search is an often-exploited feature of PubMed [1]. The relatedness between two articles is computed based on the word similarity of their articles, titles and MeSH annotations [2]. This method is very effective in finding related articles, as demonstrated by Lin et al. [1, 3]. On the average, an article A can have hundreds of related article links. The related links of A are usually ranked in order of their similarity scores.

Besides the PubMed method, co-citation analysis is another established approach for linking related information in literature [4, 5]. Existing citation based approaches mostly rely on co-citation frequency to single out related articles. For example, two articles A and B are considered to be related if they are frequently co-cited by other articles [6], or if they cite the same large set of other articles [7]. Although a large citation database is publicly available from PubMed, this useful resource has not been fully exploited for related article search.

We propose to enrich PubMed search with a new kind of related article links derived from co-citations. Instead of using citation frequency, we use citation distance (measured in the same paper) to identify related articles among co-citations. We assume that articles cited closer in the same paper would be more related. For example, articles cited in the same paragraph would be relevant to the same subtopic, thus being more related than articles cited further apart.

In this work, we focus on co-cited articles related by the possibly shortest citation distance, namely *articles cited in the same sentence*. We call these article links *Sentence Level Co-citations (SLCs)*. We show that SLCs are more related than *Paper Level Co-citations (PLCs)* (i.e. co-citations not restricted to the same sentence) in terms of various similarity metrics. We also explore another feature of SLCs. Specifically, the citing sentence of an SLC link (A, B) usually mentions the common subtopic of A and B. In many cases, the common subtopic of SLC pairs is very specific and difficult to pinpoint from simple navigation of PubMed related links.

Results

The SLCs and PLCs in our experiments have been extracted from the PMC Open Access database and from the PubMed citation database. We compare SLCs and PLCs in terms of word similarity, citation graph similarity and PubMed ranks. Word similarity is computed using the standard tf*idf model in text information retrieval [8]. Citation graph similarity is our newly defined similarity metric, which is based on findings from graph theory. Intuitively, this citation graph similarity is a probabilistic proximity between articles in their citation graph. PubMed rank of an article pair (A, B) is the rank of B among the related articles of A, provided that (A, B) is a PubMed related article link. (More details on computation are provided in Methods section).

Characterizing SLC relatedness

Word similarity of SLCs We assessed word similarity between SLCs, PLCs and 100,000 ran-

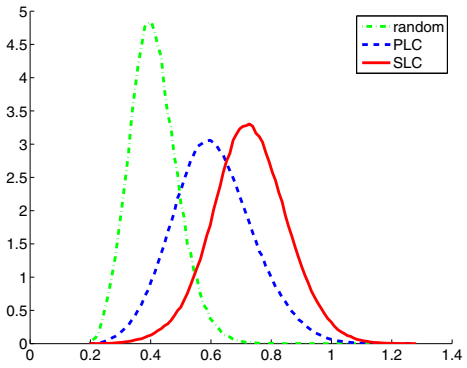

Figure 1: Distribution of word similarity among random article pairs, paper level co-cited pairs (PLC) and sentence level co-cited pairs (SLC).

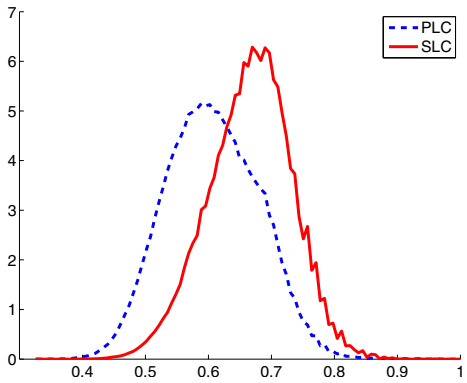

Figure 2: Distribution of citation graph similarity among paper level co-cited pairs (PLC) and sentence level co-cited pairs (SLC).

dom article pairs from PubMed (control). The empirical distributions of word similarities in different sets of article pairs are shown in Figure 1. As expected, the relatedness of PLC and SLC pairs is much higher than that of the control set. SLC pairs are (statistically) the most related in terms of word similarity. Particularly, the median similarity of SLC pairs lies around the 80^{th} percentile of the PLC pairs. The median similarity of PLC pairs is around the 99^{th} percentile of the random pairs.

Citation graph similarity of SLCs The commonly used citation similarity, namely co-citation frequency, is not useful in distinguishing SLCs from PLCs: the median co-citation frequency of PLC pairs and SLC pairs are the same (equal 1). We have defined a new citation similarity of co-cited articles, which we named *citation graph similarity*. Intuitively, this similarity is a probabilistic proximity between articles on their citation graph. We define the proximity between articles A and B to be the probability of shortest random walks making a round trip from A to B and back to A, passing through articles citing A and B in between. These random walks have the form $A \rightarrow C_1 \rightarrow B \rightarrow C_2 \rightarrow A$, where the articles C_1 and C_2 cite both A and B.

Given the new citation similarity, we can distinguish the SLC pairs from PLC pairs as in the previous section (Figure 2). Particularly, the median citation similarity of SLC pairs lies around the 80^{th} percentile of the PLC pairs.

Overlap between SLC and PubMed related article links We have seen that SLC pairs tend to have high word similarity. Since PubMed related articles are computed (effectively) using word similarity, many SLC links would be expected among the PubMed links. To estimate the overlap between SLCs and PubMed links, we randomly selected a large sample (> 5000) of SLC pairs and queried PubMed for the related article links of the articles of these pairs. Less than half of the SLC pairs were also PubMed related article links. The median PubMed rank of these SLCs is 30. This result implies that the relatedness of SLCs can be partially but not completely explained by word similarity.

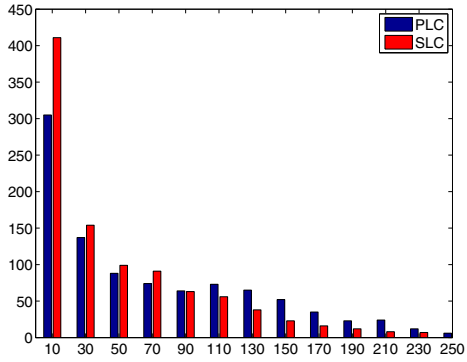

Figure 3: Histogram of PubMed ranks of 1000 SLC pairs and 1000 PLC pairs found among PubMed related article links. The median rank of SLC pairs and PLC are approximately 30 and 50, respectively.

We also carried out a similar sampling and querying procedure for PLC pairs. The comparative result is shown in Figure 3. SLC pairs tend to have higher PubMed ranking than PLCs, which confirms our finding in Figure 1.

Special features of SLC relatedness

The citing sentence S of an SLC pair (A, B) can serve as an explanation for how A and B are related. Examples of SLC pairs and citing sentences can be found in Table 1. The citing sentences describe insightful connections between the SLC articles, which can not be easily gleaned from their titles. When users navigate among PubMed related article links, they would rely

mainly on titles to select the next link. By means of SLC citing sentences, users have additional information to choose among related articles.

The examples in Table 1 give some intuition about the differences between SLC-based and word similarity based relatedness. Many of the connections between the SLC articles are derived from detailed understanding of the content of the articles. What is needed is an author who parses the two articles, and puts an explicit link (SLC) between them. This kind of high-level understanding has not been fully captured by existing similarity metrics.

Methods

Data and feature construction

We retrieved citing sentences and citation lists by processing the full-text articles provided by the PMC Open Access database. This database contains a set of about 100,000 full-text articles [9]. From this corpus, we extracted sentence level co-citations (SLCs) from the citing sentences.

Our new citation similarity is defined by random walk on a citation graph, whose edges correspond to PLC pairs. The PLCs extracted from the above mentioned 100,000 articles would form a very sparse citation graph. This graph is also inherently biased due to non-random selection of the 100,000 PMC Open Access articles. To have a more robust computation of citation graph similarity, we extracted PCLs from a larger citation database available in PubMed through the "Cited Articles" link. This PubMed database contains the citation lists of nearly a million articles. The citation graph (i.e. PLCs) of these articles would be used to computed the citation graph similarity described in the following.

Computing word similarity

Word similarity of articles can be computed using abstracts, titles and full-texts and MeSH annotations (MeSH similarity is often calculated by comparing the words of the concept descriptors). The similarity of these different types of texts can be used alone or in combination. For example, PubMed combines abstract, title and MeSH to compute word similarity, using a heuristic weighting scheme [2]. We have decided to use the simple yet effective form of word similarity, namely the word similarity computed on abstracts. The reason is that the different word similarities are correlated to some extent. Our goal is to study the trend of relatedness, thus we can sidestep the complex computational problem of selecting an optimal word similarity.

The abstracts of the all the articles are available through our MEDLINE licensed database. The abstracts have English stop words removed and are stemmed using the Lucene open-source search engine. We use the well-known tf*idf weighting scheme [8] to compute the word similarity of two abstracts A and B as

$$word\_sim(A, B) = \sum_t weight(t, A) * weight(t, B)$$

Here t is a term occurring in the articles. The function $weight(t, X)$ is the weight of the term t in an abstract X, computed as follows. Given a term t, the total number $DF(t)$ of abstracts containing t is called the *document frequency* of t. Let *Ndoc* be the total number of abstracts, then the *inverse document frequency* of t is:

$$IDF(t) = \frac{Ndoc}{DF(t)}$$

Intuitively, $IDF(t)$ measures the information content of t (i.e. rare terms would be more informative than common terms). Now let $occ(t, X)$ be the number of occurrences of term t in X, then the term frequency of t in X is:

$$TF(t) = \frac{occ(t, X)}{\sum_{u \in X} occ(u, X)}$$

That is, $TF(t)$ can be interpreted as the probability of encountering the term t in X. Finally, $weight(t, X)$ is to be proportional to the chance of seeing t in X and the information content of t:

$$weight(t, X) = TF(t, X) * IDF(t)$$

Computing citation graph similarity

Since we expect that SLC pairs have stronger semantic relatedness, we need additional features besides word similarity to characterize SLC pairs. Co-citation frequencies are not useful in distinguishing SLC pairs from PLC pairs (e.g., SLC pairs and PLC pairs have the same median 1 of co-citation frequencies).

For this purpose, we have defined a new citation similarity. Our new similarity metric is inspired by the idea of the *commute time* of random walks on graphs [10]. The commute time between two nodes A and B on a graph is the expected length of random walks on the graph going from A to B then returning to A. Commute time has been shown to be a good similarity metric on graphs and used in spectral graph clustering [11]. The computation of commute time requires the spectral decomposition of a large (albeit sparse) matrix, which is unfeasible given our computing resources. We have devised the following new similarity metric to simply approximate the commute time.

We start with two observations. First, the commute time between two articles A and B is smaller if their shorter random walks have higher probabilities. Secondly, we are mostly interested in random walks joining A and B through their citing articles. The shortest of such random walks have the form:

$$A \rightarrow C_1 \rightarrow B \rightarrow C_2 \rightarrow A$$

Rank		Title of first PMID	Title of second PMID	SLC citing sentence
1⇒2	1⇐2			
6	14	11992264: Recurrent mutation of the gene encoding sequestosome 1 (SQSTM1/p62) in Paget disease of bone.	15125799: Two novel mutations at exon 8 of the Sequestosome 1 (SQSTM1) gene in an Italian series of patients affected by Paget's disease of bone (PDB).	*Genotype-phenotype analysis has shown that SQSTM1 mutations are highly penetrant, such that between 90–100 of individuals within families who carry mutations will have developed the disease by the age of 65 years.* (from PMID 16762063)
682	1101	11779461: Wnt/Frizzled activation of Rho regulates vertebrate gastrulation and requires a novel Formin homology protein Daam1.	12533515: Coactivation of Rac and Rho by Wnt/Frizzled signaling is required for vertebrate gastrulation.	*The mechanism by which Dsh controls cell polarity and migration is unclear, but is hypothesized to involve the modulation of actin dynamics through activation of RhoA and Rac.* (from PMID 16225669)
66	91	11472631: The C-terminal domain of the Bloom syndrome DNA helicase is essential for genomic stability.	12826610: The human Bloom syndrome gene suppresses the DNA replication and repair defects of yeast dna2 mutants.	*Recombinant baculoviruses were isolated by limiting dilution and inspection by immunofluorescent microscopy to confirm nuclear localization.* (from PMID 14577841)
560	251	8692862: Skin wounds and severed nerves heal normally in mice lacking tenascin-C.	10547346: The tenascin-C knockout revisited.	*TN-C knockout mice show decreased accumulation of FN in wounded tissue and this decrease has been hypothesized to be due to disrupted incorporation of FN into the matrix in the absence of TN-C.* (from PMID 12057014)
18	28	8818392: Skin cancer prevention, early detection, and management: current beliefs and practices of Australian family physicians.	9418763: Family physicians' knowledge of malignant melanoma.	*Moreover, primary care clinicians lack confidence in their ability to diagnose melanoma.* (from PMID 10938181)
57	183	11242036: Involvement of chemokine receptors in breast cancer metastasis.	12239174: CXCR4-SDF-1 signaling is active in rhabdomyosarcoma cells and regulates locomotion, chemotaxis, and adhesion.	*Previous experimental studies have suggested that hematogenous metastasis to the lung or bone marrow is partly dependent on CXCR4 in various cancers.* (from PMID 16168106)
151	9	11238922: HIRA, the human homologue of yeast Hir1p and Hir2p, is a novel cyclin-cdk2 substrate whose expression blocks S-phase progression.	11278991: Specific phosphorylation of nucleophosmin on Thr(199) by cyclin-dependent kinase 2-cyclin E and its role in centrosome duplication.	*Decreased nuclear Cdk2 concentrations correlate with dephosphorylation of nuclear RbCdk2 has several nuclear substrates whose phosphorylation is thought to play a role in cell cycle progression.* (from PMID 11597326)
698	652	15118073: Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib.	15329413: EGF receptor gene mutations are common in lung cancers from "never smokers" and are associated with sensitivity of tumors to gefitinib and erlotinib.	*Data on the response of patients with lung cancer have demonstrated that approximately 1015% of patientwith EGFR-positive lung carcinomas have a dramatic response to EGFR tyrosine kinase inhibitors.* (from PMID 16280053)

Table 1: Examples of SLCs among PubMed related article links. The first two columns show the ranking of the second article among the related articles of the first (1⇒2), and vice versa (1⇐2). The citing sentences are verbatim.

where C_1 and C_2 are articles citing both A and B. Thus we can define the similarity between A and B to be the probability of the random walks

$$A \rightarrow C_1 \rightarrow B \rightarrow C_2 \rightarrow A \ .$$

Then the citation graph similarity of A and B is computed by:

$cite\_sim(A, B) = \sum_{C \in to(A) \cap to(B)} \frac{1}{|from(C)|} \left(\frac{1}{|to(A)|} + \frac{1}{|to(B)|} \right)$

Here, $from(X)$ and $to(X)$ denote the set of articles cited by X and the set of articles citing X, respectively.

Discussion and Conclusion

Existing approaches to enhance PubMed related article search focus mainly on reorganizations of related article links [12, 13, 14]. We propose to enrich PubMed with related articles derived from co-citation analysis. Co-citation analysis is a well-known approach for establishing relatedness between articles. It is known that more frequently co-cited articles have a higher chance to be related. However, there exists no obvious cutoff for co-citation frequency that ensures the relatedness of co-cited articles.

This work studied a particular type of article co-citations, which we call Sentence Level Co-citations (SLCs). We showed that the relatedness between articles connected by SLCs is higher than those connected by Paper Level Co-citations (PLCs). In fact, the distribution of word similarity in SLCs approximates that of word similarity among PLCs of co-citation frequency above 10 (data not shown).

We have demonstrated that SLCs capture relatedness beyond the similarity in the word content of articles. Among the PubMed related articles links, which are predominantly calculated by comparing word content, many known SLCs can not be found. Often, SLCs capture a specific subtopic between articles, and there is a need for human parsing (i.e. an author reading two abstract, and linking them with a SLC).

There are quite a few venues for future research. Let us briefly describe some possible follow-up research in the near future. First, SLCs can be considered as cited articles that are zero sentence apart. We can generalize this observation to consider cited articles that are n sentence apart. As n increases, the relatedness of cited articles decreases. The generalized SLCs would substantially expand the set of related articles. Secondly, citing sentences can provide us with a corpus of topic sentences. They can also constitute a summary of articles.

Some of the limitations of our work are as follows. First, we require full-text articles to extract SLCs. Although open access full-text articles are becoming more and more available, the reliance on full-text articles is limiting the applicability of the approach in the short term. Also, SLCs represent a sampling of related articles. The sampling is carried out by citing authors. There is certainly a large set of potential SLCs that are not or will never be cited. Nevertheless, we hope our work will stimulate more research on article relatedness that is less dependent on word similarity.

References

[1] Lin J, DiCuccio M, Grigoryan V, Wilbur WJ. Exploring the effectiveness of related article search in PubMed. TR, Uni Maryland, College Park. 2007 July;.

[2] Computation of Related Articles;. Available from: http://www.ncbi.nlm.nih.gov/entrez/query/static/computation.html.

[3] Lin J, Wilbur WJ. PubMed related articles: a probabilistic topic-based model for content similarity. BMC Bioinformatics. 2007;8(243).

[4] Synnestvedt M, Chen C. Visualizing AMIA : a medical informatics knowledge domain analysis. In: Proc AMIA Symp.; 2003. p. 1024.

[5] Synnestvedt M, Chen C, Holmes J. CiteSpace II: visualization and knowledge discovery in bibliographic databases. In: Proc AMIA Symp.; 2005. p. 724–8.

[6] Braam RR, Moed HF, van Raan AFJ. Mapping of science by combined co-citation and word analysis. I. Structural aspects. Journal of the American Society for Information Science. 1999;42(4):233–251.

[7] Tbahriti I, Chichester C, Lisacek F, Ruch P. Using argumentation to retrieve articles with similar citations: an inquiry into improving related articles search in the MEDLINE digital library. Int J Med Inform. 2006;75(6):488–495.

[8] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Information Processing & Management. 1988;24(5):513–523.

[9] PMC Open Access Subset;. Available from: http://www.pubmedcentral.nih.gov/about/openftlist.html.

[10] Lovasz L. Random walks on graphs: a survey. Combinatorics. 1996;2:353–398.

[11] Qiu H, Hancock ER. Clustering and embedding using commute times. IEEE Trans Pattern Anal Mach Intell. 2007;29(11):1873–1890.

[12] Yamamoto Y, Takagi T. Biomedical knowledge navigation by literature clustering. J of Biomedical Informatics. 2007;40(2):114–130.

[13] Plikus M, Zhang Z, Chuong CM. PubFocus: semantic MEDLINE/PubMed citations analytics through integration of controlled biomedical dictionaries and ranking algorithm. BMC Bioinformatics. 2006;7(1):424.

[14] Lu Z, Kim W, Wilbur WJ. Evaluating Relevance Ranking Strategies for MEDLINE Retrieval. J Am Med Inform Assoc. 2009;16(1):32–36.